Heart Disease

Ray Pan (yulinp3@illinois.edu)

4/29/2021

Contents

Abstract

An analysis of the heart disease data would find the best model that could help people quickly find out if they have heart disease. I trained three different machine learning algorithms to predict heart disease and have them compared. As a result, the logistic model is the most accurate when simply determine if they have the disease, and the Random Forest model is slightly better when trying to find out the number of narrowed vessels. People should use the results from the models as a reference but not a final medical decision as there's a chance of false prediction.

Introduction

Heart disease has become one of the most concerning and common causes of death in recent years. This analysis aims to find a better model that predicts the existence of heart disease in a patient by using several machine learning models. In this way, a patient would be able to find out his or her situation earlier and hopefully get an early cure out of it.

Methods

Data

I first clean up the data by creating dataset without columns containing more than 33% NAs.

```
na_prop = function(x) {
  mean(is.na(x))
}

# check proportion of NAs in each column
sapply(hd, na_prop)
```

```
##         age         sex          cp    trestbps        chol         fbs 
## 0.000000000 0.000000000 0.000000000 0.064130435 0.032608696 0.097826087 
##     restecg     thalach       exang     oldpeak       slope          ca 
## 0.002173913 0.059782609 0.059782609 0.067391304 0.335869565 0.664130435 
##        thal         num    location 
## 0.528260870 0.000000000 0.000000000
```

```
# create dataset without columns containing more than 33% NAs
hd = na.omit(hd[, !sapply(hd, na_prop) > 0.33])
```

In logistic regression, y value must be either 1 or 0, a new variable "num_log" is added to the dataset by treating v0 as no disease(0) and other values as having heart disease(1).

```
hd['num_log']<-NA
hd$num=factor(hd$num)
hd$num_log[hd$num == 'v1'|hd$num == 'v2'|hd$num == 'v3'|hd$num == 'v4'] <- 1
hd$num_log[hd$num == 'v0'] <- 0
```

Modeling

I first do a test-train split Training (70%) and Testing (30%)

```
set.seed(100)
# test-train split
hd_trn_idx = sample(nrow(hd), size = 0.7 * nrow(hd))
hd_trn = hd[hd_trn_idx, ]
hd_tst = hd[-hd_trn_idx, ]
```

Logistic

Fit a logistic regression and find out the accuracy.

```
#logistic
fit_log<-glm(num_log~.-num, data=hd_trn, family="binomial")
fit_log_prob<-predict(fit_log, hd_tst, type='response')
```

I select 0.5 as the cutoff for the positive class to distinguish True Positives, False Positives, False Negatives, True Negatives. The accuracy and no information rate of the model are printed below.

```
## [1] "Accuracy:0.837837837837838"
```

```
## [1] "No Information Rate:0.585585585585586"
```

We can see that the classifier achieves an accuracy above the no information rate.

We then find the cross-validated accuracy for the logistic regression.

```
## [1] "Cross-validated accuracy:0.789712471994025"
```

Which is still acceptable after avoiding overfitting.

A logistic model is somewhat good for a patient to quickly see whether he or she has the disease. However, it does not reflect on how many major vessels are narrowing as it assumes a patient with at least 1 major vessel with greater than 50% diameter narrowing as having heart disease. The next few models would hopefully help determine to what level they have the disease.

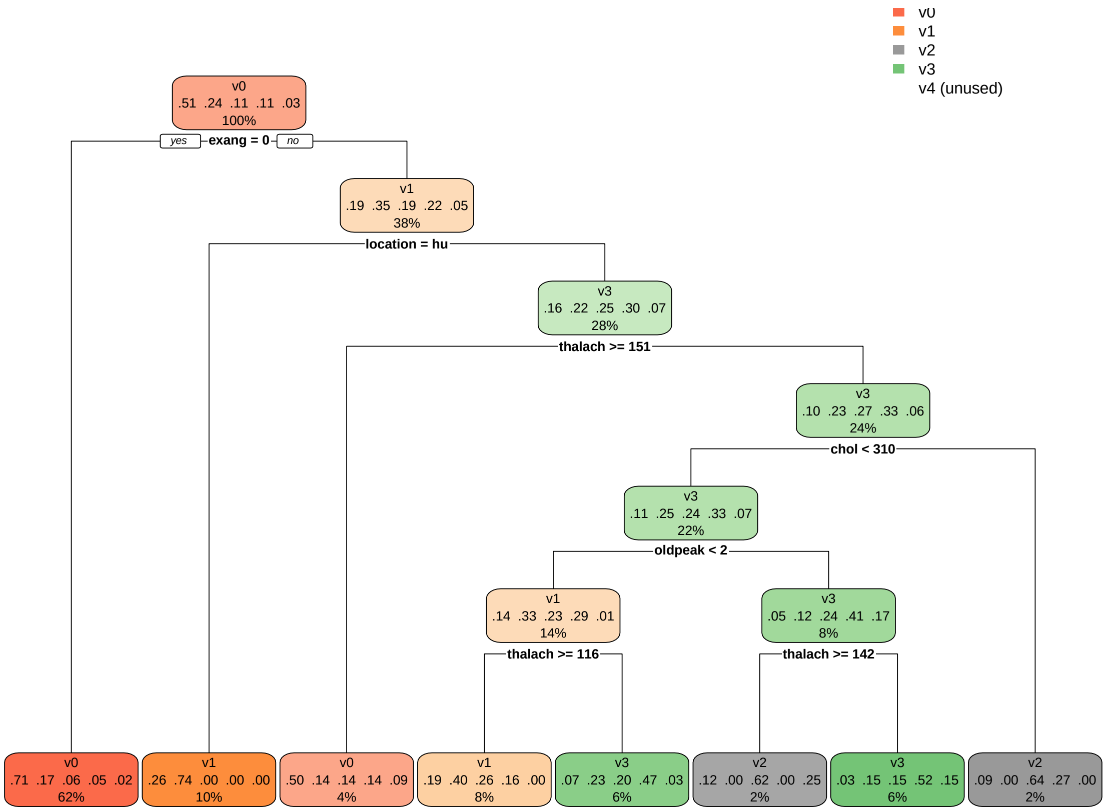
Decision Tree

I first fit a decision tree model and find the accuracy.

```
## [1] "Accuracy:0.567567567567568"
```

I then find the cross-validated accuracy for the Decision Tree model.

```
## [1] "Cross-validated accuracy:0.569697535474235"
```



With the help of plot, we could see that some important variables include exang, location, thalach, oldpeak, and chol.

It doesn't look very convincing, I would try another model.

Random Forest

The Random Forest model is used here. It combines multiple methods and would hopefully produce a better model. The accuracy and cross-validated accuracy of this model is shown below.

```
## [1] "Accuracy:0.590090090090099"
```

```
## [1] "Cross-validated accuracy:0.586696633843978"
```

We could see the results are only slightly better than the previous model.

Results

	Logistic	Decision Tree	Random Forest
Accuracy	0.8378378	0.5675676	0.5900901
Cross-Validated accuracy	0.7897125	0.5696975	0.5866966

If someone simply wants to determine whether a patient has the possibility of having heart disease, the log model would be better since we have an accuracy of 0.8378 with a sensitivity of 0.8696 and a specificity of 0.8231. While the models are good for predicting the T/F question, they might not work as expected when it is used to determine the number of narrowed major vessels. If someone wants to predict the exact number of narrowed vessels, both the decision tree and random forest models could work, but they might not be as accurate as someone needs it to be, as the slightly better model only have an accuracy of 0.59.

Discussion

While accuracy of 0.8378 is acceptable for a model, it's not as good for medical purposes. A patient should only use this model as a reference or self-check. If a severe condition occurred, the patient should directly refer to the doctor's opinion. This model is typically useful for patients who are not very concerned about their situation and would just like a quick check to save time by only providing information on important variables such as exang, location, thalach, oldpeak, and chol.

Appendix

Data Dictionary

- age - Age in years
- sex - Sex (1 = male; 0 = female)
- cp - Chest pain type (1 = typical angina; 2 = atypical angina; 3 = non-anginal pain; 4 = asymptomatic)
- trestbps - Resting blood pressure (in mm Hg on admission to the hospital)
- chol - Serum cholesterol in mg/dl
- fbs - Fasting blood sugar level > 120 mg/dl (1 = true; 0 = false)
- restecg - Resting electrocardiographic results (0 = normal; 1 = having ST-T wave abnormality: T wave inversions and/or ST elevation or depression of > 0.05 mV; 2 = showing probable or definite left ventricular hypertrophy by Estes' criteria)
- thalach - Maximum heart rate achieved
- exang - Exercise induced angina (1 = yes; 0 = no)
- oldpeak - ST depression induced by exercise relative to rest
- num - Angiographic disease status
  - `v0`: 0 major vessels with greater than 50% diameter narrowing. No presence of heart disease.

  - v1: 1 major vessels with greater than 50% diameter narrowing.
  - v2: 2 major vessels with greater than 50% diameter narrowing.
  - v3: 3 major vessels with greater than 50% diameter narrowing.
  - v4: 4 major vessels with greater than 50% diameter narrowing.

- location - location(cl = Cleveland, hu = Hungarian, ch = Switzerland, va = Virginia)

- num_log - disease status(1 = have disease, 0 = no disease)

Logistic Regression Confusion Matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction   0   1
##          0  80  24
##          1  12 106
##
##                Accuracy : 0.8378
##                  95% CI : (0.7826, 0.8838)
##     No Information Rate : 0.5856
##     P-Value [Acc > NIR] : 5.366e-16
##
##                   Kappa : 0.6721
##
##  Mcnemar's Test P-Value : 0.06675
##
##             Sensitivity : 0.8696
##             Specificity : 0.8154
##          Pos Pred Value : 0.7692
##          Neg Pred Value : 0.8983
##              Prevalence : 0.4144
##          Detection Rate : 0.3604
##    Detection Prevalence : 0.4685
##       Balanced Accuracy : 0.8425
##
##        'Positive' Class : 0
##
```

Decision Tree Confusion Matrix

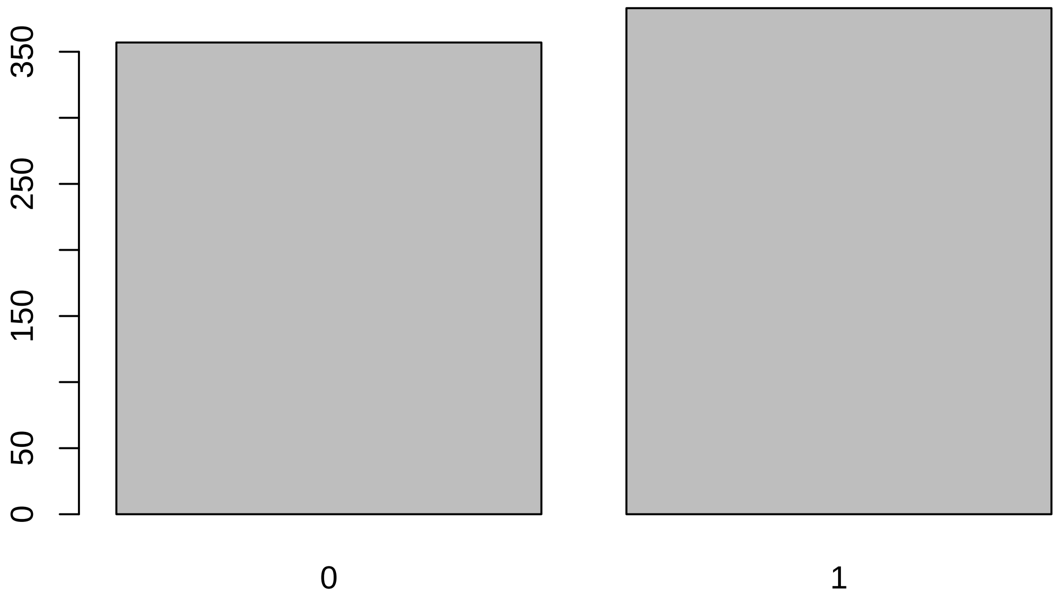
```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction v0 v1 v2 v3 v4
##         v0 80 33 10  6  3
##         v1  9 35  1  5  1
##         v2  0  2  3  0  0
##         v3  3 12  8  8  3
##         v4  0  0  0  0  0
##
## Overall Statistics
##
```

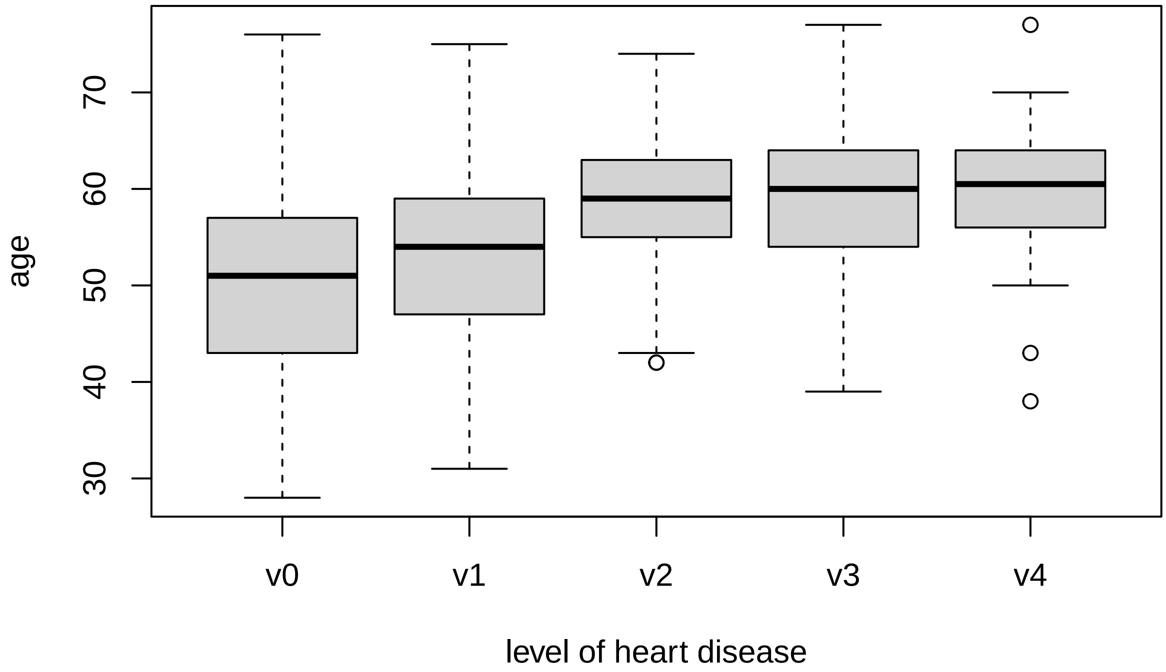
```
##                Accuracy : 0.5676
##                  95% CI : (0.4996, 0.6337)
##     No Information Rate : 0.4144
##     P-Value [Acc > NIR] : 3.065e-06
##
##                   Kappa : 0.3382
##
##  Mcnemar's Test P-Value : NA
##
## Statistics by Class:
##
##                      Class: v0 Class: v1 Class: v2 Class: v3 Class: v4
## Sensitivity             0.8696    0.4268   0.13636   0.42105   0.00000
## Specificity             0.6000    0.8857   0.99000   0.87192   1.00000
## Pos Pred Value          0.6061    0.6863   0.60000   0.23529       NaN
## Neg Pred Value          0.8667    0.7251   0.91244   0.94149   0.96847
## Prevalence              0.4144    0.3694   0.09910   0.08559   0.03153
## Detection Rate          0.3604    0.1577   0.01351   0.03604   0.00000
## Detection Prevalence    0.5946    0.2297   0.02252   0.15315   0.00000
## Balanced Accuracy       0.7348    0.6563   0.56318   0.64649   0.50000
```

Random Forest Confusion Matrix

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction v0 v1 v2 v3 v4
##         v0 84 29  5  4  1
##         v1  5 34  5  3  1
##         v2  0  9  7  6  1
##         v3  3 10  5  6  4
##         v4  0  0  0  0  0
##
## Overall Statistics
##
##                Accuracy : 0.5901
##                  95% CI : (0.5223, 0.6554)
##     No Information Rate : 0.4144
##     P-Value [Acc > NIR] : 1.021e-07
##
##                   Kappa : 0.3877
##
##  Mcnemar's Test P-Value : 0.0001785
##
## Statistics by Class:
##
##                      Class: v0 Class: v1 Class: v2 Class: v3 Class: v4
## Sensitivity             0.9130    0.4146   0.31818   0.31579   0.00000
## Specificity             0.7000    0.9000   0.92000   0.89163   1.00000
## Pos Pred Value          0.6829    0.7083   0.30435   0.21429       NaN
## Neg Pred Value          0.9192    0.7241   0.92462   0.93299   0.96847
## Prevalence              0.4144    0.3694   0.09910   0.08559   0.03153
## Detection Rate          0.3784    0.1532   0.03153   0.02703   0.00000
## Detection Prevalence    0.5541    0.2162   0.10360   0.12613   0.00000
## Balanced Accuracy       0.8065    0.6573   0.61909   0.60371   0.50000
```

Other plots




age
30
40
50
60
70
v0
v1
v2
v3
v4
level of heart disease